# Adéquation d'indices statistiques à l'interprétation de règles d'association

Hacène Cherfi, Yannick Toussaint

# Interprétation des règles d'association extraites par un processus de fouille de textes

# Interpretation of the association rules extracted by a process of text mining

Hacène Cherfi Yannick Toussaint

Équipe ORPAILLEUR
Loria - Inria Lorraine

Campus scientifique - B.P. 239
Vandœuvre-lès-Nancy F-54506 cedex
{cherfi,yannick}@loria.fr

## Résumé

*Nous proposons, dans cet article, la description d'une méthodologie d'accès et de lecture des règles d'association extraites à partir de textes. Le corpus qui a servi à notre expérience est une collection de textes sous forme de résumés d'articles scientifiques dans le domaine de la biologie moléculaire. Notre recherche porte sur : i) l'extraction des règles d'association sur des données textuelles; ii) l'association d'indices statistiques à chaque règle, ce qui permet de les ordonner; iii) l'interprétation de ces règles par un expert du domaine afin de trouver un lien entre les indices et la nature des connaissances qu'il recherche. Cet article portera essentiellement sur les deux derniers points. Nous montrons l'importance d'aider l'expert dans son interprétation des règles à l'aide des indices statistiques. Nous soulignons également la difficulté de caractériser une règle par rapport aux textes et au domaine considéré. Une discussion sur nos résultats identifie quelques points ayant un impact sur l'interprétation des règles d'association.*

## Mots-clés

Règles d'association, fouille de textes, indices statistiques, interprétation , biologie moléculaire.

## Abstract

*This paper aims at defining a methodology of access and reading of association rules extracted from texts. The corpus we used for our experiment is a collection of scientific abstracts in the field of molecular biology. Our research is related to: i) the extraction of the association rules on textual data; ii) the computation of statistical indexes for each rule, which makes it possible to order them; iii) the interpretation of these rules by an expert of the domain in order to find a link between the indexes and the nature of knowledge which he/she seeks. This article deals primarily with the last two items. We underline the importance to help the expert during this interpretation activity by using statistical indexes. We also emphasise that characterising a rule relatively to the texts and the considered domain is not an easy task. A discussion about our results highlights some points having an impact on the interpretation of the association rules.*

## Keywords

Association rules, text mining, statistical indexes, interpretation, molecular biology.

## 1 Introduction

L'utilisation de systèmes classifi catoires pour la fouille de données a été étudiée dans des domaines très divers de l'Intelligence Artifi cielle tels que : les systèmes de représentation des connaissances par objets, les logiques de descriptions, l'apprentissage, etc. Mais également dans des domaines associés comme en bases de données ou en Génie logiciel. Lors de l'utilisation d'un système classifi catoire, d'importants résultats peuvent être exploités – en dehors de la construction de classes d'individus ou de la hiérarchie d'objets. Notre article montre un aspect de l'utilisation de ces résultats pour la fouille de données dans les textes (FDT).

La fouille de texte consiste à donner à un utilisateur une vue synthétique du contenu d'une collection d'un ou de plusieurs milliers de textes (*i.e.* un corpus). La FDT exhibe des relations entre les différentes notions impliquées dans un texte ou des relations entre les textes. Ce sont des relations de généricité, de similitude, de causalité et de tendances retrouvées à partir des textes et qui sont potentiellement porteuses de sens. Les relations qui ont un sens peuvent ensuite être exploitées comme des « connaissances ».

Notre approche s'appuie, à la fois, sur des méthodes sym-

boliques et sur l'interprétation de résultats « formels ». La méthodologie générale repose sur une analyse formelle de concepts (AFC) [9, 21] appliquée aux textes.
Nous cherchons à collecter les relations entre textes par le biais de *règles d'association*. Ces règles constituent des indicateurs précieux pour la veille technologique et pour l'analyse de l'information. Les règles d'association extraites sont destinées à être lues, interprétées et évaluées par un utilisateur qui, dans notre cas, est un expert du domaine. Néanmoins, leur nombre est élevé; nous cherchons donc à les ordonner en associant des indices statistiques destinés à classer les règles entre elles.
Parallèlement, nous demandons à l'expert d'évaluer chacune des règles par rapport à son intérêt dans le domaine et de faire un classement de la règle la plus interprétable à la règle la moins interprétable. L'étude des valeurs des différents indices des règles nous permet de voir si certains reflètent l'intérêt que l'expert porte à certaines règles [4]. La confrontation des résultats formels (calcul des règles d'association, calculs des indices) à la réalité du domaine (l'appréciation de l'expert) est inédite. Afi n de faciliter l'interprétation par l'expert, nous défi nissons un environnement de navigation dans l'ensemble des règles.
La section (2) donne la défi nition d'une règle d'association et introduit les indices statistiques associés. Dans la section (3) nous décrivons les données, les deux expérimentations que nous avons menées et nous caractérisons, dans l'ensemble, les règles extraites. Nous détaillons, en section (4), l'interprétation des règles par l'expert. Les sections (5) et (6) font un bilan de l'expérimentation et un parallèle avec des travaux similaires.

# 2 Fouille de textes

Notre processus de fouille est fondé sur l'utilisation de méthodes symboliques. C'est la combinaison :

(I) - d'une méthode formelle d'extraction de règles d'association;

(II) - d'un classement des règles suivant des indices statistiques;

(III) - d'un mécanisme interactif d'accès aux règles et au contenu des textes;

(IV) - d'une interprétation par l'expert.

Les concepts formels, tels qu'ils sont défi nis en AFC, permettent l'extraction des règles d'association (I) à travers la construction des ensembles fermés « fréquents » générés par l'algorithme *Close* [15]. Les indices statistiques calculés en (II) sont des mesures de pondération affectées à chaque règle. Ces indices donnent un poids à chaque règle et permettent alors de faire des "classements" de règles. Tous les indices introduits (voir section 2.2) ne s'avéreront pas utilisables pour notre processus de fouille de textes.
L'environnement de navigation hypertexte (III) aide l'expert du domaine à accéder aux règles d'association, aux indices et aux textes qui ont permis l'extraction de la règle. La lecture des règles permet de les interpréter (IV) par rapport au domaine considéré.

## 2.1 Définition d'une règle d'association

Une règle d'association [11, 14, 1] est du type :

$$R \ : \ t_1 \wedge t_2 \Longrightarrow t_3 \wedge t_4 \wedge t_5 \qquad (1)$$

(où $\mathtt{t_1}$ ... $\mathtt{t_5}$ sont des termes). Elle est constituée d'une conjonction de termes en partie gauche (que nous appellerons B) impliquant une conjonction de termes en partie droite (notée H). La règle sera donc notée $\mathtt{R} : \mathtt{B} \Longrightarrow \mathtt{H}$.
Les règles d'association ont été, initialement, utilisées pour trouver des régularités, des corrélations pour la « fouille » dans des bases de données relationnelles de grande taille. Elles ont été, plus tard, appliquées à la fouille de textes [7]. L'explication intuitive de la règle (1) est que tous les documents qui sont caractérisés par les termes $\{\mathtt{t_1},\mathtt{t_2}\}$ sont aussi caractérisés par $\{\mathtt{t_3},\mathtt{t_4},\mathtt{t_5}\}$.
À l'origine, deux principaux indices ont été associés aux règles : le *support* et la *confiance*.

**Définition du support.** Le support d'une règle d'association est l'ensemble des textes participant à son extraction (*i.e.* sa validité). Il représente le nombre de textes qui sont décrits par les termes présents en partie gauche et droite de la règle (on dira par la suite : le nombre de *documents* qui *vérifient* B et H). Pour la règle (1) le support est :

$$sup\,[B \Longrightarrow H] = \ nombre\ de\ docs\ v\acute{e}rifiant\ \{t_1,t_2,t_3,t_4,t_5\} \qquad (2)$$

Le support peut également être exprimé relativement au nombre total de document du corpus. C'est la probabilité d'apparition de l'ensemble des documents correspondant à $\mathtt{B} \wedge \mathtt{H}$ et que nous noterons par la suite $\mathtt{P(B\ ,\ H)}$ :

$$P(B,\,H) = \frac{sup\,[B \Longrightarrow H]}{nombre\ total\ de\ documents} \in [0,\,1] \qquad (3)$$

**Définition de la confiance.** La confi ance mesure le degré de validité d'une règle, c'est-à-dire lorsqu'il existe des contre-exemples de documents qui vérifi ent B mais pas nécessairement tous les termes de H. Pour la règle (1), la confi ance vaut

$$conf = \frac{nombre\ de\ documents\ v\acute{e}rifiant\{t_1,t_2,t_3,t_4,t_5\}}{nombre\ de\ documents\ v\acute{e}rifiant\{t_1,t_2\}} \qquad (4)$$

En termes probabilistes, la confi ance mesure la probabilité conditionnelle de H sachant B : $\mathtt{conf}\ [\mathtt{B} \Longrightarrow \mathtt{H}] = \mathtt{P(H \mid B)}$. Lorsque la confi ance vaut 1, la règle est dite ***totale***. Dans le cas contraire, la règle est dite ***partielle*** à **x** %. Afi n de réduire le nombre des règle à extraire, ces processus utilisent des valeurs seuils `minsup` et `minconf` qui sont, respectivement, la valeur minimum souhaitée pour le *support* et la *confiance*.

## 2.2 Indices associés aux règles d'association

Le support et la confi ance ne permettent pas, à eux seuls, d'indiquer la « qualité » d'une règle par rapport au domaine considéré. Nous introduisons d'autres indices statistiques qui apportent des informations supplémentaires et permettent différents classements des règles.

Les termes rares ne sont pas présents dans les règles car nous fixons un seuil `minsup` en dessous duquel la règle correspondante n'est pas extraite. Lorsque les termes ont une fréquence dans le corpus qui est proche du seuil, les règles correspondantes ne sont pas celles qu'on est tenté d'analyser en premier. En revanche, des termes trop répandus dans le corpus n'apportent pas d'information « particulière » puisque tout terme du corpus impliquera un terme fréquent parce qu'ils apparaissent ensemble dans beaucoup de textes. Cette différence d'apparition des termes a conduit à la définition d'autres indices dont certains ont un comportement symétrique pour B et H. Le sens de l'implication sous-jacente à une règle d'association n'est pas reflétée par un $\texttt{indice}_i$ tel que : $\texttt{indice}_i[\texttt{B} \Longrightarrow \texttt{H}] = \texttt{indice}_i[\texttt{H} \Longrightarrow \texttt{B}]$. À l'inverse, d'autres indices ne sont pas symétriques.

**L'intérêt.** Si B et H sont indépendants (*i.e.* $\texttt{P(B,H)} = \texttt{P(B)} \times \texttt{P(H)}$), la confiance (4) vaut $\texttt{P(H)}$. Une règle d'association extraite à partir de termes indépendants n'est pas intéressante. L'*intérêt* est défini par :

$$int\,[B \implies H] = \frac{P(B,H)}{P(B) \times P(H)} \qquad (5)$$

L'intérêt mesure la déviation de l'indépendance entre B et H. Cet indice favorise le classement des règles ayant des termes rares aux dépens des règles ayant des termes trop répandus dans le corpus.

**La conviction.** L'indice (5) est totalement symétrique, il ne reflète pas l'implication $\texttt{B} \Longrightarrow \texttt{H}$. La *conviction* proposée par [3] est :

$$conv\,[B \implies H] = \frac{P(B) \times P(\neg H)}{P(B,\neg H)} \qquad (6)$$

La conviction mesure également la déviation de l'indépendance mais pour les contre-exemples $\texttt{B} \wedge \neg\texttt{H}$[1]. La conviction a l'avantage de souligner le caractère implicatif de $B$ vers $H$ (*i.e.* en estompant le côté symétrique de l'indice (5)). Cet indice n'est applicable que pour les règles *partielles* car $\texttt{P(B, }\neg\texttt{H)}$ vaut 0 pour des règles *totales* et de ce fait $\texttt{conv}\,[\texttt{B} \Longrightarrow \texttt{H}]$ est une valeur non calculable.

**La dépendance.** L'indice de dépendance permet de calculer l'apport de B dans la règle. Cet indice est classiquement utilisé en probabilités. La *dépendance* est définie par :

$$dep\,[B \implies H] = \mid P(H \mid B) - P(H) \mid \qquad (7)$$

Plus la différence entre la probabilité d'avoir H sachant B et la probabilité d'avoir H est grande, moins H et B sont liés. La règle bien classée par l'indice de dépendance s'en trouve plus *renforcée*. Deux autres indices liés à (7) permettent de corriger la différence de représentativité entre B et H.

1. $\texttt{B} \Longrightarrow \texttt{H}$ est logiquement équivalent à $\neg(\texttt{B} \wedge \neg\texttt{H})$

Le premier, appelé *nouveauté*, mesure la différence de la probabilité d'avoir B et H et la probabilité que les parties B et H soient indépendantes.

$$nov\,[B \Longrightarrow H] = P(H,B) - P(B) \times P(H) \qquad (8)$$

Cet indice se ré-écrit en : $\texttt{dep}\,[\texttt{B} \implies \texttt{H}] \times \texttt{P(B)}$. La nouveauté atténue la dépendance des règles dont la partie B est trop répandue dans le corpus.

Le second appelé : *satisfaction*, normalise la dépendance entre B et $\neg$H par $\texttt{P}(\neg\texttt{H})$.

$$sat\,[B \implies H] = \frac{(P(\neg H) - P(\neg H \mid B))}{P(\neg H)} \qquad (9)$$

la *satisfaction* s'écrit aussi : $\texttt{sat}\,[\texttt{B} \Longrightarrow \texttt{H}] = \frac{\texttt{dep}\,[\texttt{B} \Longrightarrow \neg\texttt{H}]}{\texttt{P}(\neg\texttt{H})}$ ou bien $(1 - \texttt{int}\,[\texttt{B} \Longrightarrow \neg\texttt{H}])$. La *satisfaction* atténue la dépendance lorsque la partie H est trop répandue dans le corpus et favorise le classement des règles par rapport à la dépendance des parties B et H lorsque les termes qui y apparaissent sont rares dans le corpus. Cet indice n'est significatif que pour les règles *partielles*. Pour les règles totales $\texttt{sat}\,[\texttt{B} \Longrightarrow \texttt{H}]$ vaut toujours 1.

**L'étonnement.** Cet indice, présenté dans [13], est défini pour mesurer l'*affirmation* : différence entre la *confirmation* $\texttt{P(B, H)}$ et l'*infirmation* $\texttt{P(B, }\neg\texttt{H)}$ d'une règle. Cet indice permet de rechercher les règles dites « étonnantes ». Plus H est rare dans le corpus, plus il est étonnant de trouver une bonne affirmation de la règle. Les auteurs remarquent que les règles dont l'indice est supérieur à un certain seuil sont insensibles au bruit, c'est-à-dire aux données non désirées liées à des biais dans l'indexation dans notre cas. L'*étonnement* (appelé aussi *surprise*) s'écrit :

$$spr\,[B \implies H] = \frac{(P(B,H) - P(B,\neg H))}{P(H)} \qquad (10)$$

Comme (9), cet indice n'est valable que pour les règles *partielles* à cause de la présence de $\texttt{P(B, }\neg\texttt{H)}$.

# 3 Expérimentations

Dans cette section, nous décrivons le corpus de données, les deux expérimentations puis les résultats auxquels nous avons abouti.

## 3.1 Description des données

Notre corpus est constitué de 1 407 documents d'environ 200 000 mots, soit environ 6 Mø. Un *document* est constitué d'un *identifiant* unique (*i.e.* un numéro), d'un titre, d'un (ou des) auteur(s), du résumé sous forme textuelle et d'une liste de termes caractérisant ce résumé. Les textes sont en anglais et traitent de la biologie moléculaire, plus particulièrement de la mutation de gènes provoquant une résistance aux antibiotiques. Le corpus nous a été fourni par

l'INIST[2]. La figure (fig. 1) donne l'exemple du document numéro 000391 de notre corpus.
L'indexation des textes s'appuie sur la notion de **terme** (*i.e.* mot ou ensemble composé de mots). L'indexation par les termes est plus appropriée que l'indexation par les mots simples pour caractériser un texte et rendre compte de son contenu. Comme le souligne [5] : « *Les termes permettent généralement de limiter l'ambiguïté et d'augmenter la précision* ». Non seulement grâce au repérage de notions mieux dénommées, mais également grâce au réseau constitué par la terminologie. Pour faire cette indexation, nous avons opté pour l'outil FASTR [12]. C'est un analyseur syntaxique fondé sur les grammaires d'unification [16] et, plus précisément, sur la forme logique des Grammaires d'Arbres Adjoints [20]. FASTR recherche, dans des séquences textuelles acceptables, le maximum de termes qui s'y trouvent par identification de ces termes à partir d'une *liste contrôlée* (appelée nomenclature terminologique). Les formes variantes de termes reconnues dans les textes sont ramenées à leur terme préférentiel. Par exemple, on voudrait que le terme "*transfer of capsular biosynthesis genes*" indexe le texte par son terme préférentiel "*gene transfer*".

Document 000391
Titre : Sequencing of gyrase and topoisomerase IV quinolone-resistance-determining regions of Chlamydia trachomatis and characterization of quinolone-resistant mutants obtained In vitro.

Auteur(s) : Dessus-Babus-S; Bebear-CM; Charron-A; Bebear-C; de-Barbeyrac-B

Texte : The L2 reference strain of Chlamydia trachomatis was exposed to sub-inhibitory concentrations of ofbxacin (0.5 microg/ml) and sparfbxacin (0.015 microg/ml) to select fluoroquinolone-resistant mutants. In this study, two resistant strains were isolated after four rounds of selection. The C. trachomatis mutants presented with high-level resistance to various fluoroquinolones, particularly to sparfbxacin, for which a 1,000-fold increase in the MICs for the mutant strains compared to the MIC for the susceptible strain was found. The MICs of unrelated antibiotics (doxycycline and erythromycin) for the mutant strains were identical to those for the reference strain. The gyrase (gyrA, gyrB) and topoisomerase IV (parC, parE) genes of the susceptible and resistant strains of C. trachomatis were partially sequenced. A point mutation was found in the gyrA quinolone-resistance-determining region (QRDR) of both resistant strains, leading to a Ser83–>Ile substitution (Escherichia coli numbering) in the corresponding protein. The gyrB, parC, and parE QRDRs of the resistant strains were identical to those of the reference strain. These results suggest that in C. trachomatis, DNA gyrase is the primary target of ofbxacin and sparfbxacin.

Mot(s)-clé(s) : "characterization" "determine region" "dna" "doxycycline" "erythromycin" "escherichia coli" "gyrA gene" "gyrase" "gyrB gene" "mutation" "ofbxacin" "parC gene" "parE gene" "point mutation" "protein" "quinolone" "sequencing" "sparfbxacin" "substitution" "susceptible strain" "topoisomerase"

FIG. 1 – *Document* 000391 *: composé d'un titre, des auteurs, du texte complet et des mots-clés*

## 3.2 Description des expériences

Deux expériences ont été menées avec ce corpus :

– La première fut réalisée avec une indexation entièrement automatisée. Il en a résulté que l'ensemble des documents a été indexé par un total de 22 885 termes qui correspondent à 3 337 termes différents, avec une moyenne de 16,26 termes par document. Parmi ces termes, 1 762 (soit 52,8 %) étaient des termes n'apparaissant qu'une seule fois en index (*i.e.* termes *hapax*) et beaucoup de bruit lié au découpage des termes en sous-termes par FASTR. Cette distribution des termes dans le corpus, considéré comme un « éparpillement » de l'information, est un biais bien connu en analyse de l'information textuelle. Il est dû, notamment, au bruit lié aux termes, périphériques du domaine, utilisés par les auteurs des textes;

– La seconde expérience avec des termes filtrés à la main par les documentalistes de l'INIST. Ce filtrage manuel permet d'éliminer une grande partie du bruit. Il résulte que l'ensemble des documents a été indexé par un total de 14 374 termes pour 1 361 documents (le reste des documents n'ayant plus de termes d'indexation). Ces termes correspondent à 632 termes différents (soit 18,94 % du nombre de termes différents de la 1ère expérience) , avec une moyenne de 10,56 termes par document. À noter, qu'il n'y a pas de termes apparaissant moins de 5 fois dans l'indexation et 49 % des termes apparaissent entre 5 et 15 fois.

L'ensemble des couples ("document" - "liste de termes") constituent les tableaux de données en entrée pour les deux expériences.

## 3.3 Résultats obtenus

Pour notre première expérience, les règles *totales* extraites de support minimal 10 étaient au nombre de 1 202 dont 713 étaient des règles avec un support $\in [10,15]$ (soit 59,31 % des règles obtenues). Lorsqu'on fixe le support à 1, nous obtenons plus de 460 000 règles. Cette première expérience nous a permis de tester la robustesse de nos calculs sur un corpus de taille moyenne. Ces paramètres d'exécution machine sont très peu pénalisants en termes d'occupation mémoire et de temps machine d'exécution. Les calculs ne durent pas plus de 1 à 2 minutes. Cette expérience a également montré les inconvénients d'une indexation entièrement automatique. Les règles obtenues n'étaient pas interprétables et beaucoup trop nombreuses. Comme le souligne [10] : « *... le nombre de règles calculé peut être très élevé et les tâches de dépouillement, d'interprétation et de synthèse des résultats peuvent alors devenir extrêmement complexes, voire inextricables, pour l'utilisateur* ».
Dans la 2ème expérience, les règles *totales* extraites de support minimal 10 étaient au nombre de 128. Lorsqu'on ramène le support à 1, nous augmentons sensiblement leur nombre puisque nous obtenons 163 175 règles.

# 4 Interprétation par l'expert

Nous avons soumis les 128 règles obtenues lors de la seconde expérience à un expert-documentaliste de l'INIST. Les règles n'ont pas été classées pour laisser à l'expert une libre appréciation. Il est important, pour nous, de repérer quelles règles lui paraissent "interprétables". Puis nous avons confronté ces règles aux indices calculés pour chacune d'elles. Les règles ont été, pour la plupart, interprétées.
Avant de décrire les règles, nous rappelons quelques no-

2. INstitut de l'Information Scientifique et Technique

tions utiles à la compréhension des règles à interpréter.
L'information génétique est contenue dans l'ADN présent dans chacune des cellules qui composent tout organisme vivant. L'ADN détermine le cycle de vie, de division, et les autres caractéristiques de ces cellules. Cette information est présente sous forme de gènes qui codent pour des protéines [22].
Les antibiotiques permettent d'inhiber la synthèse protéique, ils stoppent la fabrication d'enzymes par la bactérie (celle-ci ne se reproduit plus et/ou la cellule meurt). Mais les gènes de la bactérie mutent et l'antibiotique ne peut plus reconnaître la bactérie cible et s'y fixer. C'est le schéma général du phénomène de résistance aux antibiotiques.
Ce qui suit est une sélection de certaines règles interprétées par l'expert.

## 4.1 Les meilleures règles interprétées

Nous commençons par commenter la règle qui reflète une description du domaine d'activité.

Numéro :000045
Règle : "determine region" "gyrA gene" "gyrase" "mutation" $\Longrightarrow$ "quinolone"
Support : "11"
Documents participant à la règle : "000391" "000395" "000491" "000616" "000650" "000781" "001126" "001149" "001186" "001196" "001317"
Intérêt : "17.012" Dépendance : "0.941"

Cette règle indique que les 11 documents cités décrivent la mutation du gène "gyrA" qui code pour l'enzyme "gyrase" dans un fragment ou une zone précise de l'ADN. Cet enzyme est responsable de la résistance aux antibiotiques de la famille des "Quinolones". Pour avoir le schéma complet du mécanisme de résistance, il manque le nom de la bactérie, qui n'est pas le même pour les 11 documents (pour le document 000391 (cf. fig. 1) , il s'agit de "Chlamydia trachomatis" alors que pour le document 000491 c'est "Pseudomonas aeruginosa", etc.).

Numéro :000114
Règle : "parC gene" "sequence" $\Longrightarrow$ "gyrA gene"
Support : "11"
Documents participant à la règle : "000104" "000346" "000545" "000619" "000776" "000781" "001035" "001037" "001186" "001265" "001348"
Intérêt : "21.603" Dépendance : "0.954"

Cette règle fait ressortir le fait que le gène "parC" a été découvert plus récemment que le gène "gyrA". Mais que ces deux gènes sont liés dans leurs mutations (par mutation combinée). Chaque fois qu'on parle de "parC", les auteurs font référence aussi à "gyrA".

Numéro :000011
Règle : "bla gene" "escherichia coli" $\Longrightarrow$ "lactamase"
Support : "12"
Documents participant à la règle : "000672" "000768" "000797" "000843" "000958" "000963" "000983" "000989" "001175" "001244" "001290" "001311"
Intérêt : "10.007" Dépendance : "0.900"

La bactérie "Escherichia Coli" possède le gène "bla" donc fabrique l'enzyme $\beta$-Lactamase et sera donc résistante à la famille des $\beta$-lactams.

Numéro :000095
Règle : "mecA" "meticillin" $\Longrightarrow$ "mecA gene" "staphylococcus aureus"
Support : "12"
Documents participant à la règle : "000095" "000678" "000744" "000812" "000875" "000886" "001087" "001090" "001111" "001168" "001239" "001324"
Intérêt : "80.059" Dépendance : "0.988"

Grâce à l'utilisation de la "Meticillin", on inhibe le gène "mecA" et on détruit la "Staphylococcus Aureus". Cette bactérie chez l'homme est à l'origine d'un problème de santé publique grave car elle est responsable de milliers de morts dans le monde.
Cette règle est à rapprocher par le contenu à la règle suivante :

Numéro :000077
Règle : "grlA gene" "mutation" $\Longrightarrow$ "staphylococcus aureus"
Support : "11"
Documents participant à la règle : "000067" "000309" "000312" "000347" "000562" "000767" "000786" "000874" "001039" "001059" "001246"
Intérêt : "7.561" Dépendance : "0.868"

Elle dit simplement que la résistance à la "Staphylococcus Aureus" est liée à la mutation du gène "grlA".

## 4.2 Les règles indésirables

Nous allons caractériser les règles que l'expert a jugé "inutiles", et néanmoins vraies, selon trois critères qui sont : le bruit lié à l'indexation par FASTR, le bruit lié à la synonymie et le bruit lié aux énumérations de termes.

**Bruit lié à l'indexation.** L'analyseur FASTR, dans son processus d'extraction de termes, procède par reconnaissance de termes les plus longs puis par découpage en sous-termes. Trois règles illustrent ce biais.

Numéro : 000108
Règle : "mycobacterium tuberculosis" $\Longrightarrow$ "tuberculosis"
Support : "72"
Documents participant à la règle : "000039" "000063" "000083" "000086" "000091" "000137" "000147" "000165" "000194" "000197" "000199" "000216" "000217" "000262" "000277" "000278" "000294" "000300" "000306" "000327" "000399" "000408" "000427" "000429" "000470" "000472" "000473" "000474" "000488" "000492" "000515" "000519" "000528" "000540" "000551" "000557" "000563" "000567" "000582" "000626" "000667" "000687" "000728" "000748" "000769" "000804" "000809" "000815" "000821" "000849" "000913" "000945" "000969" "000991" "000993" "001003" "001014" "001049" "001069" "001085" "001102" "001180" "001191" "001201" "001207" "001238" "001245" "001259" "001284" "001300" "001301" "001302"
Intérêt : "14.956" Dépendance : "0.933"

Dans cette règle à très fort support de 72 documents, on voit que le re-découpage de "Mycobacterium Tuberculosis" en "Tuberculosis" est généré par FASTR. Le terme "Tuberculosis" n'empêche pas l'interprétation de cette règle car la "Tuberculosis" (*tuberculose*) reste cohérente avec la bactérie "Mycobacterium Tuberculosis" qui la provoque. Les choses sont différentes pour la règle suivante :

Numéro : 000012
Règle : "broad host range" $\Longrightarrow$ "host range"
Support : "11"
Documents participant à la règle : "000026" "000049" "000111" "000156" "000249" "000337" "000452" "000549" "000723" "000763" "000940"
Intérêt : "80.059" Dépendance : "0.988"

Le terme "broad host range" en B a un sens particulier dans ce domaine mais pas le terme de la règle en H, où le qualificatif "broad" a été éliminé du terme.

Enfi n la règle suivante n'a pas lieu d'être :

Numéro : 000087
Règle : "infection" "urinary infection" $\Longrightarrow$ "urinary tract"
Support : "10"
Documents participant à la règle :"000037" "000124" "000313" "000768" "000781" "000896" "000981" "000983" "001012" "001056"
Intérêt : "104.692" Dépendance : "0.990"

L'unique terme d'index exact serait : "urinary tract infection".

**Bruit lié à la synonymie.** Malgré l'utilisation d'une nomenclature terminologique par FASTR (cf. (3.1)), tous les synonymes n'ont pas été ramenés au terme préférentiel de la nomenclature du fait du manque d'exhaustivité de cette nomenclature. Les deux règles suivantes en sont une illustration :

Numéro : 000073
Règle : "epidemic strain" $\Longrightarrow$ "outbreak"
Support : "16"
Documents participant à la règle : "000060" "000196" "000255" "000270" "000324" "000388" "000414" "000510" "000767" "000787" "000844" "000854" "001022" "001088" "001221" "001260"
Intérêt : "17.449" Dépendance : "0.943"

Numéro : 000127
Règle : "topoisomerase" $\Longrightarrow$ "gyrase"
Support : "15"
Documents participant à la règle : "000074" "000080" "000212" "000256" "000347" "000391" "000491" "000712" "000874" "000877" "000892" "000920" "001186" "001265" "001317"
Intérêt : "30.932" Dépendance : "0.968"

Les termes en partie gauche et droite sont des synonymes. Le fort support de ces règles par rapport aux précédentes souligne le fait que les biologistes manipulent dans leurs textes indifféremment un terme ou son synonyme. C'est particulièrement gênant pour des analyses automatiques de texte ou pour la fouille de textes.

**Bruit lié aux énumérations.** Certains gènes sont systématiquement associés à d'autres gènes. La raison n'est pas liée au fait que l'article s'intéresse à tous les gènes cités dans le texte. Il peut s'agir seulement de situer un gène par rapport à des gènes analogues ou des sous-unités de gènes qui ont un comportement similaire.

Numéro : 000048
Règle : "determine region" "gyrA gene" "parE gene" $\Longrightarrow$ "parC gene" "quinolone"
Support : "10"
Documents participant à la règle : "000104" "000391" "000545" "000619" "001032" "001035" "001126" "001186" "001196" "001317"
Intérêt : "45.367" Dépendance : "0.978"

Certains textes parmi les 10 ont étudié la sous-unité "E" du gène "par", mais citent les autres noms de gènes "gyrA" et "parC" pour situer ou défi nir le gène "*parE*".

**Choix entre deux règles proches.** Alors qu'en analyse formelle de concepts, nous recherchons des règles décrivant des concepts « génériques » avec fort *support*, nous avons été étonnés de voir qu'entre les règles 000006 et 000005 suivantes :

Numéro : 000006
Règle : "aztreonam" "enzyme" $\Longrightarrow$ "lactamase"
Support : "16"
Documents participant à la règle : "000664" "000672" "000779" "000787" "000839" "000843" "000923" "000963" "000998" "001034" "001054" "001101" "001130" "001176" "001290" "001295"
Intérêt : "10.007" Dépendance : "0.900"

Numéro : 000005
Règle : "aztreonam" "clavulanic acid" "enzyme" $\Longrightarrow$ "lactamase"
Support : "11"
Documents participant à la règle : "000672" "000787" "000839" "000843" "000963" "001034" "001054" "001101" "001130" "001176" "001290"
Intérêt : "10.007" Dépendance : "0.900"

l'expert ait préféré la seconde règle car le nom de l'acide aminé "Clavunalic" qui inhibe l'enzyme $\beta$-Lactamase y est cité. Pourtant la première règle possède un *support* beaucoup plus fort.

# 5 Bilan de l'expérimentation

Dans cette section, nous allons mettre en évidence des liens entre les règles extraites présentées en (4) et les différents indices des sections (2.1) et (2.2). Par la suite, nous ferons quelques remarques sur cette confrontation entre les indices et les règles interprétées par l'expert.

## 5.1 Adéquation entre les règles présentées et les indices

Nous nous sommes intéressés aux règles listées en (4), que l'expert a interprétées, pour essayer de les confronter aux valeurs des indices. Tout d'abord, il faut noter que les règles présentées à l'expert étaient des règles *totales*. Elles ont donc toutes une *confiance* de 100 %. Les règles présentées, ont majoritairement des *supports* assez proches compris entre 10 et 12. C'est d'ailleurs une caractéristique globale de toutes les règles extraites sauf celles dues au bruit lié à la synonymie (cf. section (4.2), 2$^{\text{ème}}$ sous-section). En général, nous constatons qu'un fort support ne signifie pas nécessairement un intérêt de l'expert pour la règle (cf. section (3.2), première expérience).
En ce qui concerne l'*intérêt*, on vérifie bien qu'il caractérise une différence de représentativité des termes en B et H. La règle 000087 : ("infection" "urinary infection" $\Longrightarrow$ "urinary tract"), d'*intérêt* : 104.692, possède le terme "infection" qui domine par sa fréquence dans cette règle. Il apparaît 180 fois dans le corpus, contre seulement 13 fois pour "urinary tract" et 12 fois pour "urinary infection". Pour la règle 000095 : ("meca" "meticillin" $\Longrightarrow$ "meca gene" "staphylococcus aureus"), d'*intérêt* : 80.059, le déséquilibre est plutôt constaté du côté de H : le terme "Staphylococcus Aureus" est présent 180 fois dans le corpus alors que les termes "mecA gene" et "meticillin" apparaissent beaucoup moins: 18 et 52 fois respectivement. La règle 000012 : ("broad host range" $\Longrightarrow$ "host range"), ayant un *intérêt* de 80.059, illustre la rareté de B et de H en même temps : "host range" apparaît 17 fois et le nombre d'apparition de "broad host range" dans le corpus est de seulement 11 fois. Les indices d'*intérêt* de ces règles sont les valeurs les plus grandes pour l'ensemble des règles.
En ce qui concerne la *dépendance*, nous avons constaté que

les règles les plus dépendantes étaient celles relatives au biais dû à l'indexation dans les règles : 000087 et 000012, citées plus haut, avec respectivement 99 % et 98 %. Les règles découlant du biais introduit par la synonymie ont également un fort indice de *dépendance*, la règle 000127 : ("topoisomerase" $\Longrightarrow$ "gyrase") a un indice de 96 % et la règle 000073 : ("epidemic strain" $\Longrightarrow$ "outbreak") a un indice de 94 %.
L'indice *nouveauté* le plus élevé est celui de la règle 000108 ("mycobacterium tuberculosis" $\Longrightarrow$ "tuberculosis"), il vaut 4.9 %. Cet indice apporte plus d'informations pour les règles *partielles*. C'est également le cas des indices de *satisfaction* et d'*étonnement* (cf. 2.2).

### 5.2 Éléments de discussion

Nous donnons ici quelques éléments de réflexion qui ont suivi notre expérimentation. Notre approche s'appuie sur une description booléenne (présence *vs.* absence) des termes dans les documents. Cette représentation ne prend pas en compte la fréquence d'apparition des termes à l'intérieur du document et/ou plus globalement dans l'ensemble des documents. L'approche est différente en Recherche d'Information où l'on associe à un terme une *pondération*. Cette pondération est fonction de la fréquence par rapport aux autres termes dans le document et de la fréquence dans la globalité du corpus. Cela permet de faire un classement entre documents contenant les mêmes termes en réponse à une requête de l'utilisateur. Notre méthode paraît en ce sens plus sensible à la phase d'indexation. Si un terme est absent de l'indexation (*i.e.* silence), cela peut entraîner la disparition d'une règle du fait des seuils de *support* et de *confiance* choisis.
Bien que le corpus soit spécialisé (résistance des microbes aux antibiotiques), nous constatons une assez grande disparité des termes retenus à l'indexation. On retrouve ce phénomène régulièrement en analyse automatique de corpus. Comme nous l'avions souligné en (3.2), un texte intégral fait souvent référence à divers termes périphériques au domaine considéré qui introduisent du bruit.
Enfin, nous remarquons que l'*implication* dans une règle ne porte pas d'information particulière pour le jugement de sa qualité par l'expert. Peut-être que le réflexe d'indexation lié au statut de documentaliste de notre expert fait qu'il voit les règles comme une liste de termes. Si nous lui présentions des règles ayant un minimum de termes en B et un maximum en H, peut-être que l'expert verrait un sens à l'*implication*. Nous n'avons pas trouvé ce genre de règle, mais en prenant un *support* plus faible de 5, nous obtenons la règle suivante :

Règle : "Sequence" "parE gene" $\Longrightarrow$ "Quinolone" "Resistance" "determine region" "gyrA gene" "parC gene" "parE gene"
Support : "5"

Cela est à rapprocher du phénomène de disparité des termes souligné précédemment.
Un autre point de discussion concerne l'utilité, voire la nécessité, d'une navigation hypertextuelle dans les règles. Cette navigation correspond au besoin de l'expert de pouvoir accéder rapidement au contenu des textes auxquels font référence les documents associés à une règle à interpréter. Cela dénote, parfois, l'ambiguïté des termes en B et en H. Pour réaliser ce travail, les règles ont été codées dans une structure XML. Les règles, les valeurs d'indices et les documents associés sont alors visualisables grâce à un navigateur Web (cf. voir un aperçu en fig. 2). L'utilisateur accède aux titres de tous les documents vérifiant la règle en cliquant sur le numéro de règle. Il accède à l'aperçu (voir fig. 1) en cliquant sur un numéro de document. Un classement des règles par indice est également disponible.

## 6 Approches comparables

Certains travaux se sont intéressés à l'extraction efficace des règles d'association par une structuration préalable des données dans un espace de généralisation [2], d'autres comme [17] retrouve les règles à partir de la construction explicite du système classificatoire - ici, il s'agissait d'un treillis de concepts - et de l'organisation des relations d'héritage entre les concepts du treillis. Tout concept formel (ou nœud) du treillis permet d'extraire une règle de la forme :

$$t_i \Longrightarrow TH \tag{11}$$

où les $t_i \in \{t_1,...,t_p\}$ constituent l'ensemble TP des termes *propres* d'un concept et TH constitue l'ensemble des termes *hérités* (*i.e.* termes appartenant aux concepts « pères » du nœud courant). Mais ces méthodes demeurent liées à la gestion, très coûteuse en espace mémoire et en temps de calcul, d'une structure de données en amont (espace de généralisation et treillis de concepts) que nous n'exploitons pas dans notre processus.
Dans les travaux de [6], on part de schémas de sous-catégorisation pour « apprendre » une hiérarchie de concepts (*i.e.* ontologie) par une classification hiérarchique ascendante (CHA) et par l'utilisation de relations grammaticales dans les textes, par exemple :

$$[Secher]\ COD\ < aliment > \tag{12}$$

$$[Secher]\ CC\ < air > \tag{13}$$

Ces schémas sont appris à partir d'exemples contenus dans un corpus étiqueté sur les recettes de cuisine. Toutes les occurrences du verbe "sécher" font apparaître un aliment en complément d'objet direct et un terme comme "air" en complément circonstanciel. [18] reprend le corpus étiqueté par les schémas de sous-catégorisation et cherche à trouver les dépendances les plus pertinentes entre des concepts et des ensembles de documents en donnant une mesure d'intensité aux règles d'association extraites. L'intensité dans les règles d'association est également utilisée dans [10] par le calcul d'une pondération des règles avec une fonction entropique tenant compte, à la fois des contre-exemples à la règle et à sa contraposée $\neg H \Longrightarrow \neg B$.
Enfin, dans [8], la recherche de règles se fait en typant B et H sur des termes filtrés, cela permet de descendre jusqu'à

des indices de *confiance* très faibles (de l'ordre de 0.1) . Par exemple, chercher tous les établissements industriels qui ont fait alliance ou qui ont fusionné : "intuit corp" "novell corp" $\Longrightarrow$ "merger".

# 7 Conclusion et perspectives

L'extraction de règles d'association est souvent exploitée dans le processus de fouille de données et de textes. Cependant, l'interprétation de ces règles et l'évaluation de leur qualité aussi bien par des indices statistiques que par des experts du domaine restent difficiles à maîtriser. Le nombre de règles extrait ne permet pas l'exploitation « efficace » des régularités et d'éventuelles "connaissances" qui émergent d'un grand corpus de textes, car le texte est différent des données que l'on traite classiquement en fouille de donées.

Nous avons combiné l'utilisation d'indices statistiques avec une approche symbolique. L'objectif étant de sélectionner les règles à analyser en vue de les présenter à un expert du domaine pour leur interprétation. Cette sélection permet de réduire le nombre de règles à analyser. Nous nous sommes intéressés à l'extraction des règles dites *totales*. Nous avons trouvé que deux de ces indices : l'*intérêt* et la *dépendance* correspondent à des règles que l'expert a jugé *pertinentes*. Les autres indices cités dans l'article ne sont applicables et/ou significatifs que pour les règles dites *partielles*.

Une autre aide consiste à réaliser une interface Web pour la navigation entre règles, indices et documents correspondants.

Il est nécessaire de poursuivre ce travail par l'extraction des règles *partielles* pour ce même corpus, afin de tester l'apport des autres indices : *conviction*, *nouveauté*, *satisfaction* et *étonnement*. La réduction du bruit lié à la synonymie peut être fait en combinant la nomenclature déjà utilisée durant l'indexation automatique par un thésaurus plus complet comme par exemple : une partie du métathésaurus de l'UMLS [19]. Enfin, nous comptons tester l'apport d'autres indices introduits dans la littérature (gain d'entropie, gini, laplace, etc.).


## Remerciements

Nous tenons à remercier M. Alain Zasadzinski, expert documentaliste à l'INIST pour l'interprétation des règles d'association obtenues par notre processus de fouille de textes. Nous remercions également M. Jean Royauté de nous avoir fourni le corpus indexé ainsi que la région Lorraine pour sa participation financière à nos travaux de recherche.


# Références


[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*, pages 478–499, Santiago, Chile, september 1994. Extended version: IBM Research Report RJ 9839.

[2] I. Bournaud and M. Courtine. Un Espace de Généralisation pour l'Extraction de Règles d'Association. In H. Briand and F. Guillet, editors, *Actes EGC'01 : Journées Extraction et Gestion des Connaissances*, volume 1 of *1-2*, pages 129–140, Nantes, France, January 2001. Éditions Hermès.

[3] S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In *Proceedings of the ACM SIGMOD'97 Conference on Management of Data*, volume 36, Tucson, USA, May 1997.

[4] H. Cherfi and Y. Toussaint. Extraction et Interprétation des Règles d'association pour la Fouille de Textes. In *Actes de l'Atelier A3CTE-01 : Applications, Apprentissage, Acquisition des connaissances à partir de textes électroniques*, pages 15–16, Grenoble, Juin 2001. Plate-forme AFIA. Résumé (Version courte).

[5] N. Faraj, R. Godin, R. Missaoui, S. David, and P. Plante. Analyse d'une méthode d'indexation automatique basée sur une analyse syntaxique de texte. *Canadian Journal of Information and Library Science / Revue l'information et la bibliothéconomie*, 21(1):1–21, 1996.

[6] D. Faure, C. Nédellec, and C. Rouveirol. Acquisition of Semantic Knowledge using Machine learning methods: The System ASIUM. Technical Report ICS-TR-88-16, LRI Université Paris-Sud, Janvier 1998.

[7] R. Feldman and I. Dagan. Knowledge Discovery in Textual Databases (KDT). In U. M. Fayyad and R. Uthurusamy, editors, *Proceedings of the 1st International Conference on Data Mining and Knowledge Discovery*, Montreal, CA, August 1995. AAI Press.

[8] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir. Text mining at the term level. *Lecture Notes in Artificial Intelligence: Principles of Data Mining and Knowledge Discovery*, 1510(1):65–73, September 1998.

[9] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, 2000.

[10] R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. In H. Briand and F. Guillet, editors, *Actes EGC'01 : Journées Extraction et Gestion des Connaissances*, volume 1 of *1-2*, pages 69–80, Nantes, France, January 2001. Éditions Hermès.

[11] J.L. Guigues and V. Duquenne. Familles minimales d'implication informatives résultant d'un tableau de données binaires. *Mathématiques, Informatique et Sciences Humaines*, 95:5–18, 1986.

[12] C. Jacquemin. FASTR : A Unification-Based Front-End to Automatic Indexing. In *Proceedings of Information Multimedia Information Retrieval Systems and Management*, pages 34–47, New-York, October 1994. Rockfeller University.

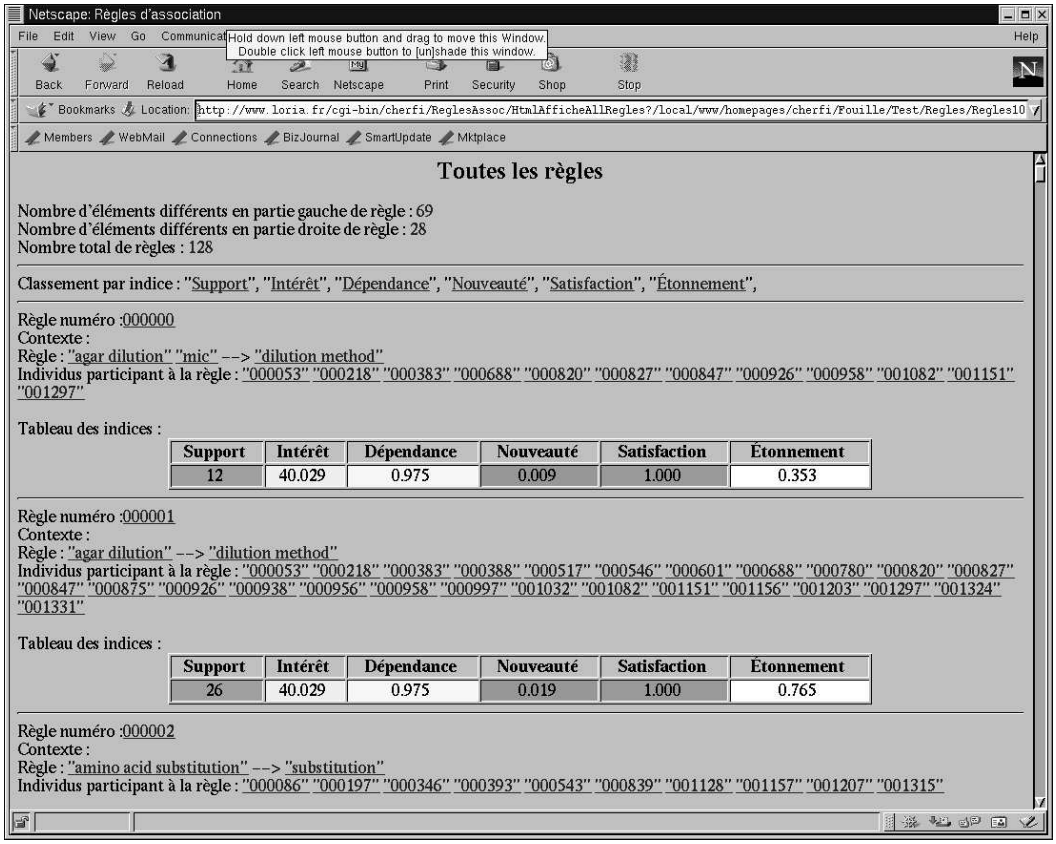


FIG. 2 – *Aperçu de l'interface de navigation Web*


[13] Y. Kodratoff and J. Azé. Rating the Interest of Rules Induced from Data and from Texts. À paraître, 2002.

[14] M. Luxenburger. Implications partielles dans un contexte. *Mathématiques, Informatique et Sciences Humaines*, 29(113):35–55, 1991.

[15] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46, 1999.

[16] S. M. Shieber. *An Introduction to Unification-Based Approaches to Grammar*. Center for the Study of Language and Information, Stanford University, Stanford, CA, 1986.

[17] A. Simon. *Outils classificatoires par objets pour l'extraction de connaissances dans les bases de donnés*. PhD thesis, Université Henri Poincaré - Nancy 1, Nancy, France, Septembre 2000.

[18] E. Suzuki and Y. Kodratoff. Discovery of Surprising Exception Rules based on Intensity of Implication. In *Proc. of the 2nd Eur. Symp. on Principles of Data Mining and Knowledge Discovery PKDD'98*, pages 10–18, Nantes, France, September 1998.

[19] UMLS. The Unified Medical Language System. 11th edition, National Library of Medicine, 2000.

[20] K. Vijay-Shankar. Using descriptions of trees in a tree-adjoining grammar. *Computational Linguistics*, 18:481–518, 1992.

[21] R. Wille. Restructuring lattices theory: an approach based on hierarchies of concepts. In I. Rival, editor, *Ordered Sets*, pages 445–470, Banff, Ca., 1982. NATO advanced institute on ordered sets, D. Reidel, Dordrecht-Boston.

[22] J. Zaccai and C. Garrec. *Les macromolécules du vivant - Structure, dynamique et fonctions*. Éditions CNRS, Paris, 1998.